Luca L 10/3/89

Editorial for American Journal of Human Genetics - draft.

How can one study individual variation for three billion nucleotides of human genome?

by L. Luca Cavalli-Sforza.
Genetics Dept.
Stanford University

In addition to the investigation of transmission, mutation, and action of genes, the study of individual variation is one of the main aims of genetics. Without a study of both normal and pathological variation, the harvest to be expected from the gigantic effort of sequencing the three billion nucleotides forming the human genome would not be fully reaped. The choice of an individual to be sequenced has been the subject of some debate and even the basis for jokes. The question of individual variation is the obvious source of this preoccupation. It is not important which individual is sequenced for a given region, but it is clearly important that some insight into individual variation be acquired. Unfortunately, the effort of producing the whole sequence just once is so great, that at the moment the idea of sequencing even two individuals instead of one cannot be seriously entertained. On the other hand, knowledge of the evolutionary conservation of DNA segments is of considerable importance. Highly variable segments are often junk DNA. There are, however, interesting

exceptions to this rule. The higher variability of some segments which are certainly not junk, e.g. in the HLA super-super gene, or in variable regions of immunoglobulin genes is more than an object of biological curiosity, and there are other similar examples beyond the immune system. An intron such as that in an globin gene may be highly conserved, and demands an explanation, evolutionary or not.

How can one resolve this impasse without waiting indefinitely for more formidable techniques? One classical strategy offers a simple solution. It is the same used, for instance, by the government of China for taking the 1982 population census without analysing one billion individuals: take a sample. China is not the only instance of census by sample, but it is my favored example because in a recent collaboration with professor Du Ruofu, of the Genetics Institute of Academia Sinica, we have been able to obtain 540,000 Chinese surnames from a stratified random sample of 1/2000-th of the Chinese population.

If one tested the variation of 100 individuals for a number of DNA segments representing globally one ten thousandth of the whole genome, the total effort of sequencing the whole genome would be increased by one per cent with respect to that necessary for sequencing a single individual. I believe the results would be sufficiently rewarding to justify a greater effort, such as that of

analyzing a sample of one thousandth of the genome. This would bring the cost of estimating individual variability up to 10% of the total effort. But one could start with a less ambitious program and amplify it later, if results justified it. These calculations are based on the reasonable assumption that costs are approximately proportional to the number of sequenced nucleotides.

The number of individuals, and the segments to be chosen for study should be considered with care. Let us take for simplicity an average length of 300 nucleotides for the DNA segments; the number of possible segments is then 10 million. One thousand of these would be chosen for the minimal effort scheme of 1% increase in global cost. The choice of a length of 300 bp was suggested by the fact that, even today, this segment length is easily amenable to PCR amplification followed by direct sequencing. With 1000 segments one would have already a reasonable number of segments from each of the various possible categories: exons, introns, promoters, enhancers, any other region potentially involved in regulation, repeated sequences (both transcribed and non transcribed), and others. Each category could be given a specific weight for the purpose of forming a stratified random sample.

If one chose to sequence less than 100 individuals, the fraction of segments to be studied could increase

proportionately. Is it necessary to examine so many individuals? I believe it is, but the number 100 is, of course, flexible. We know that in highly conserved regions, the probability that two random chromosomes differ at one nucleotide is less than one in a thousand, while in highly variable regions it may be as low as one in a hundred, but these are orders of magnitude. Our current knowledge is limited, however, and affected by sequencing errors. 100 individuals are obviously too few to be certain that a particular nucleotide is highly conserved; one would need more than 1000 individuals for this purpose, a realistic sample size only under special circumstances. With 100 individuals, the probability of detecting variation in a sequence of 10 nucleotides is 2/3 even with fairly high conservation. More sophisticated estimates and statistical procedures could be developed. Here I am only interested in indicating orders of magnitude.

Among other general considerations, an obvious one is the choice of the sample of individuals. "Immortalized" lymphocyte cultures are preferable for ensuring a sufficient amount of DNA, and for other possible tests. Cultures happen to be available for some representative aboriginal populations from very different parts of the world (12 at the moment) in a collection started in 1984 through a collaboration between Stanford (A. Bowcock, J. Hebert, A.

Lin and myself) and Yale (J. and K. Kidd). They have been immortalized for studying DNA polymorphisms from an evolutionary point of view (see LCS et al. 1985, Cold Spring Harbor Symp.; A. Bowcock et al. 1988, Gene Geography for first results; observations on the first hundred polymorphisms are summarized in a MS ready to be sent to press). Incidentally, we have many requests for DNA samples from colleagues who are aware of this little publicized effort. We were forced to ration our positive responses in order not to stop our regular research entirely, but are currently working out solutions for meeting these demands more satisfactorily. Using cultures such as these for generating the 100 or so individuals wanted for studying individual variation the chance of finding variants would be increased compared with that expected if a geographically more restricted population is studied. Nevertheless, the major attraction of this strategy would be that of generating a body of data which could enormously increase our understanding of human evolution.

However, this sample is not entirely ideal. If one found in an individual a variant of, for example, the TATA box of a known gene one might want to study the regulation of that gene in that individual, and possibly in his or her family. Many subjects from our collection would be extremely difficult to study, since they in most cases live in utterly

remote locations. In many cases the analysis of gene expression would be limited to that in lymphocytes. If this aspect is deemed to be important, it may be better to set up a new collection of individuals, still chosen from a variety of ethnic groups, but making sure that they are easily available and willing to cooperate in further studies. One might also compromise by using a mixed sample, partly of representative aboriginal populations from remote areas, and partly of more easily accessible individuals.

These considerations seem timely, and others may want to reanalyse and extend them. Clearly the human genome project can be much more rewarding if appropriate attention is given to human variability. Some might find "normal" individual variation less important than "pathological" variation, but the boundary between the two is often difficult to define. The human genome project may be very useful for helping to find the specific sequences responsible for certain genetic diseases, in conjunction with linkage analysis on pedigrees of given affections. But many other genes, which do not directly cause disease, are potentially important for medicine. For instance, the metabolism of drugs is subject to considerable genetic variation, which would certainly be classified in the "normal" range. The same could be said of genes involved in hormone action, genes for growth factors etc. Thus,

scientific curiosity is far from being the only motive for studying the so called "normal" individual variation.

Needless to say, similar strategies for the study of variation can be applied to other organisms, which are also part of the sequencing program.

8/89